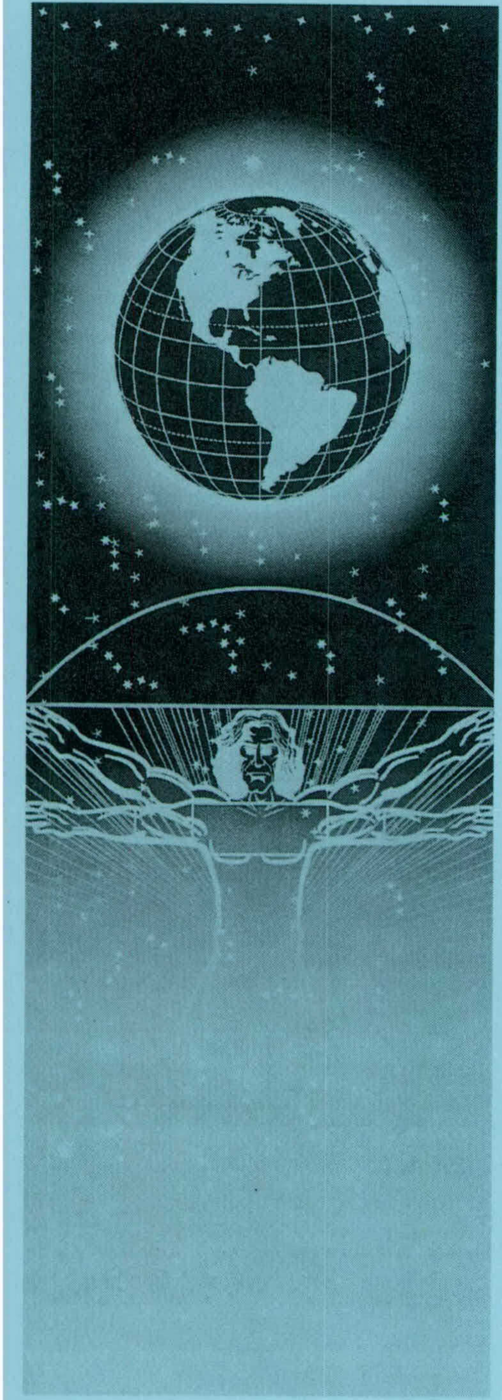
AFRL-HE-WP-TR-2004-0164



UNITED STATES AIR FORCE RESEARCH LABORATORY

AFRL/HECP SPEAKER RECOGNITION SYSTEMS FOR THE 2004 NIST SPEAKER RECOGNITION EVALUATION

RAYMOND E. SLYH
ERIC G. HANSEN
TIMOTHY R. ANDERSON

DECEMBER 2004

FINAL REPORT FOR THE PERIOD OCTOBER 1999 TO SEPTEMBER 2004

Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
2255 H Street
Wright-Patterson AFB OH 45433-7022

NOTICES

TECHNICAL REVIEW AND APPROVAL

AFRL-HE-WP-TR-2004-0164

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

//Signed//

MARIS M. VIKMANIS
Chief, Warfighter Interface Division
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* 31-12-2004	**2. REPORT TYPE** Final Report	**3. DATES COVERED** *(From - To)* October 1999 - September 2004

4. TITLE AND SUBTITLE

AFRL/HECP Speaker Recognition Systems for the 2004 NIST Speaker Recognition Evaluation

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

62202F

6. AUTHOR(S)

Slyh, Raymond E.
Hansen, Eric G.
Anderson, Timothy R.

5d. PROJECT NUMBER

7184

5e. TASK NUMBER

10

5f. WORK UNIT NUMBER

03

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory, Human Effectiveness Directorate
Warfighter Interface Division, Collaborative Interfaces Branch
Air Force Materiel Command
Wright-Patterson AFB OH 45433-7022

8. PERFORMING ORGANIZATION REPORT NUMBER

AFRL-HE-WP-TR-2004-0164

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory, Human Effectiveness Directorate
Warfighter Interface Division
Collaborative Interfaces Branch
Air Force Materiel Command
Wright-Patterson AFB OH 45433-7022

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

AFRL-HE-WP-TR-2004-0164

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The National Institute of Standards and Technology sponsors a yearly evaluation of speaker recognition systems. This report summarizes the systems submitted by AFRL/HECP for the 2004 Speaker Recognition Evaluation. The evaluation consisted of seven training conditions by four testing conditions for a total of 28 task conditions. AFRL/HECP submitted systems for all 28 task conditions, one of only three groups to do so out of the 24 participating groups from twelve countries. In addition, AFRL/HECP submitted unsupervised adaptation systems for four of the conditions. A total of ten different systems were submitted across the conditions (not counting unsupervised adaptation modes). These systems were various combinations of the scores from eight component systems. Component systems unique to AFRL/HECP's submission included a system based on glottal model parameters; a system based on formant center frequencies, formant bandwidths, and fundamental frequency; and a system based on mel-frequency cepstral coefficients (MFCCs) and phoneme-specific Gaussian mixture models (GMMs).

15. SUBJECT TERMS

speaker recognition, glottal model, formants, cepstral coefficients, Gaussian mixture model, score fusion

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Raymond E. Slyh
Unclassified	Unclassified	Unclassified	Unlimited	30	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-9248

This page intentionally left blank.

PREFACE

Since 1996, the National Institute of Standards and Technology has sponsored a yearly evaluation of speaker recognition systems. AFRL/HEC has participated in these evaluations since 2001. This report summarizes the systems submitted by AFRL/HECP for the 2004 Speaker Recognition Evaluation. The evaluation consisted of seven training conditions by four testing conditions for a total of 28 task conditions. In addition, for each standard (nonadaptive) system submitted for a given task condition, participants could submit the same system operated in an unsupervised adaptation mode. AFRL/HECP submitted systems for all 28 task conditions, one of only three groups to do so out of the 24 participating groups from twelve countries. In addition, AFRL/HECP submitted unsupervised adaptation systems for four of the conditions. A total of ten different systems were submitted across the conditions (not counting unsupervised adaptation modes). These systems were various combinations of the scores from eight component systems. Component systems unique to AFRL/HECP's submission included a system based on glottal model parameters; a system based on formant center frequencies, formant bandwidths, and fundamental frequency; and a system based on mel-frequency cepstral coefficients (MFCCs) and phoneme-specific Gaussian mixture models (GMMs). These systems were developed over the course of Workunit 71841003, Robust Voice Processing and Identification.

The authors would like to acknowledge support from the following groups: (1) MIT Lincoln Laboratory for their MFCC/GMM system and the LNKnet neural network package; (2) Prof. Bryan Pellom of the University of Colorado at Boulder for the Sonic speech recognizer, acoustic models, and a phoneme "language" model; (3) Cambridge University for their Statistical Language Modeling Toolkit (originally developed at Carnegie Mellon University) and their Hidden Markov Model Toolkit (HTK); and (4) KTH for their Wavesurfer audio analysis tool and the Snack toolkit.

Further support for this work was provided by General Dynamics Advanced Information Systems (GDAIS) under contract F41624-97-D-6004 and by Sytronics, Incorporated under contract F33615-98-D-6000. Mr. Brian Ore of GDAIS developed the segmentation system described here and a speech activity detector based on HTK. He also developed code to compute cepstral coefficients from linear prediction coefficients and did some work in combining the system scores. Mr. Anthony Halley of GDAIS generated glottal model features for the evaluation data and generated phoneme labels using Sonic. Dr. Vince Schmidt of Sytronics ported the original parallel glottal model code of the first author to the Parallel Virtual Machine (PVM) library and made other improvements resulting in faster parallel execution.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Since 1996, the National Institute of Standards and Technology (NIST) has sponsored a yearly evaluation of speaker recognition systems.[1] AFRL/HEC has participated in these evaluations since 2001. This report summarizes the systems submitted by AFRL/HECP for the 2004 Speaker Recognition Evaluation. The evaluation consisted of seven training conditions by four testing conditions for a total of 28 task conditions. In addition, for each standard (nonadaptive) system submitted for a given task condition, participants could submit the same system operated in an unsupervised adaptation mode. AFRL/HECP submitted systems for all 28 task conditions, one of only three groups to do so out of the 24 participating groups from twelve countries. In addition, AFRL/HECP submitted unsupervised adaptation systems for four of the conditions. A total of ten different systems were submitted across the conditions (not counting unsupervised adaptation modes). These systems were various combinations of the scores from eight component systems. Component systems unique to AFRL/HECP's submission included a system based on glottal model (GM) parameters; a system based on formant center frequencies, formant bandwidths, and fundamental frequency (FMBWF0); and a system based on mel-frequency cepstral coefficients (MFCCs) with phoneme-specific Gaussian mixture models (GMMs). These unique systems were developed over the course of Workunit 71841003, Robust Voice Processing and Identification. This report describes the component and combined systems and shows the performance of these various systems for some of the task conditions of the 2004 and (for comparison purposes) 2003 evaluations.

An outline of this report is as follows. The next chapter briefly describes the evaluation framework. Chapter 3 describes the various component systems as well as a speech activity detector (SAD) based on a hidden Markov model (HMM), which was used with some of the GMM-based component systems. The chapter also discusses the unsupervised adaptation strategy and the speaker segmentation system used for the evaluation. Chapter 4 discusses the submitted systems. Chapter 5 presents some results for the various systems in the form of detection error tradeoff (DET) plots, and Chapter 6 presents some conclusions and suggestions for future work.

[1] See http://www.nist.gov/speech/tests/spk/index.htm for the various evaluation plans

2 EVALUATION OVERVIEW

The NIST 2004 Speaker Recognition Evaluation consisted of 28 task conditions that were a combination of seven training conditions and four testing conditions [1]. The training and testing conditions involved various amounts of data and were defined in terms of a conversation side—namely, the last five minutes from a six-minute conversation by two strangers generally based on a single given topic. The training conditions were defined by the following training data:

10 sec: An excerpt from a single channel conversation side estimated to contain approximately 10 seconds of speech;

30 sec: An excerpt from a single channel conversation side estimated to contain approximately 30 seconds of speech;

1 side: A single channel conversation side of approximately five minutes total duration;

3 sides: Three single channel conversation sides involving the same speaker;

8 sides: Eight single channel conversation sides involving the same speaker;

16 sides: Sixteen single channel conversation sides involving the same speaker; and

3 convs: Three summed-channel conversations, formed by sample-by-sample summing of the two sides of actual conversations, each including a common speaker (the target of interest) and a second speaker not participating in the other two conversations.

The testing conditions were defined by the following testing data:

10 sec: An excerpt from a single channel conversation side estimated to contain approximately 10 seconds of speech;

30 sec: An excerpt from a single channel conversation side estimated to contain approximately 30 seconds of speech;

1 side: A single channel conversation side of approximately five minutes total duration; and

1 conv: A summed channel conversation, formed by sample-by-sample summing of the two sides of an actual conversation.

All of the training data were collected over telephone channels including cellular and land line handsets. The testing data were mostly telephone speech but included some data from various types of microphones. Most of the training and testing data were in English, but some conversations involving bilingual speakeres were collected in Arabic, Mandarin, Russian, and Spanish.

Word transcripts for all English-language training and testing segments were provided to the evaluation participants by NIST. The transcripts were generated by BBN using a speech recognizer based on their conversational-telephone-speech recognizer submitted for the NIST 2003 Rich Transcription Evaluation (RT-03). It is estimated that the word error rate of the system was in the 20–30% range.

3 COMPONENT SYSTEMS

Table 1 shows the component systems run for the various training and testing conditions. Five of the systems were based on Gaussian mixture models (GMMs) and two were based on language modeling. The FMBWF0 system was based on the first three formant center frequencies, the first three formant bandwidths, and the fundamental frequency, F0. The GM system was based on parameters from a glottal model. The LPCC system was based on cepstral coefficients and delta cepstral coefficients computed from the linear prediction polynomial determined in a closed-phase analysis of each speech signal. The MFCC system was based on mel-frequency cepstral coefficients and delta cepstral coefficients. The PS-MFCC system was similar to the MFCC system except that separate GMMs were computed for each phoneme for each speaker. The PLM system was based on applying language modeling techniques to phoneme labels determined by a phoneme recognizer, while the WLM system was based on applying language modeling techniques to the words as determined by a speech recognition system. Systems listed with "(UA)" were run in nonadaptive mode and also with unsupervised adaptation (UA).

3.1 GMM-Based Systems

The GMM-based systems all used Version 2.1 of the MIT Lincoln Laboratory (MIT-LL) system [2]. In general, only frames labeled as speech by the MIT-LL *xtalkN* energy-based speech activity detector (SAD) were used. The GMMs used diagonal covariance matrices, and the number of mixtures depended on the feature set. Background, target, and T-norm models [3] used: (1) 512 mixtures for the GM and FMBWF0 systems, (2) 2048 mixtures for the MFCC and LPCC systems, and (3) 1024 or 2048 mixtures depending on the phoneme for the PS-MFCC system. In adapting a target or T-norm model from the background model, different strategies were used depending on the feature set. The weights, means, and variances were adapted for the GM and FMBWF0 systems, while only the means were adapted for the MFCC, PS-MFCC, and LPCC systems. These adaptation strategies were determined to be the best for their respective feature sets based on experiments using data from NIST evaluations of prior years. The data used to build background and T-norm models came from databases available from the Linguistic Data Consortium.[2]

The background models for the FMBWF0, GM, LPCC, and MFCC systems were gender-independent, using a total of 13 hours of data from 74 male and 74 female speakers taken from Switchboard II Phase 3 and Switchboard Cellular I. The channel mix was as follows: (1) 40 speakers from cellular channels (mostly GSM and unknown cellular, with only a few CDMA) and (2) the rest evenly split between electret and carbon button land line handsets.

The background models for the PS-MFCC, PLM, and WLM systems were independent of gender, using approximately 42 hours of data from 250 male and 250 female speakers taken from Switchboard II Phases 2 and 3. The data for these background models only included land line (mostly electret handset) data originally used in the NIST 2002 and 2003 Extended Data Tasks for splits 6–10.

For most systems, a score normalization technique known as T-norm was applied [3]. Let $S(U, C)$ be the score from some system for a test utterance, U, against a claimant model, C. Let $\{T_1, \ldots, T_N\}$ be a set of N (T-norm) speakers not found in the background model set and not found in the the NIST 2004 evaluation set. One tests U against each of the N T-norm models and computes the mean, μ_U, and standard deviation, σ_U, of the scores from the T-norm

[2] See http://www.ldc.upenn.edu

Table 1: Component Systems Versus Training and Testing Condition

Training Condition	Testing Condition			
	10 sec	30 sec	1 side	1 conv
10 sec	MFCCs, FMBWF0, LPCCs, GM	MFCCs, FMBWF0, LPCCs, GM	MFCCs, FMBWF0, LPCCs	MFCCs with HMM SAD & clustering
30 sec	MFCCs, FMBWF0, LPCCs, GM	MFCCs, FMBWF0, LPCCs, GM	MFCCs, FMBWF0, LPCCs	MFCCs with HMM SAD & clustering
1 side	MFCCs, FMBWF0, LPCCs, GM	MFCCs, FMBWF0, LPCCs, GM	MFCCs, PS-MFCCs, LPCCs, FMBWF0 (UA), WLM (UA), PLM (UA)	MFCCs with HMM SAD & clustering
3 sides	MFCCs, FMBWF0	MFCCs, FMBWF0	MFCCs, PS-MFCCs, FMBWF0 (UA), WLM (UA), PLM (UA)	MFCCs with HMM SAD & clustering
8 sides	MFCCs, FMBWF0	MFCCs, FMBWF0	MFCCs, PS-MFCCs, FMBWF0 (UA), WLM (UA), PLM (UA)	MFCCs with HMM SAD & clustering
16 sides	MFCCs, FMBWF0	MFCCs, FMBWF0	MFCCs, PS-MFCCs, FMBWF0 (UA), WLM (UA), PLM (UA)	MFCCs with HMM SAD & clustering
3 conv	MFCCs with HMM SAD & clustering	MFCCs with HMM SAD & clustering	MFCCs with HMM SAD & clustering	MFCCs with HMM SAD & clustering

models. The adjusted score, $S_T(U,C)$, (after applying T-norm) for U against model C for the system is:

$$S_T(U,C) = \frac{S(U,C) - \mu_U}{\sigma_U},$$

provided that $\sigma_U \neq 0$.

For the FMBWF0, GM, LPCC, and MFCC systems, the T-norm was gender-independent using a total of nine hours of data from 50 male and 50 female speakers from Switchboard II Phase 3 and Switchboard Cellular I. The channel mix was as follows: 50 speakers from cellular channels (mostly GSM and unknown cellular, with only a few CDMA) and the rest split between electret and carbon button land line handsets (68% electret and 32% carbon button). For the 10-second and 30-second test cases, the T-norm models were built from the first 30 seconds of data from the original set of T-norm models.

For the PS-MFCC and WLM systems, the T-norm was gender-independent using approximately 17 hours of data from 50 male and 50 female speakers from Switchboard II Phases 2 and 3, all outside the scope of the NIST 2003 Extended Data Task. Each model was built using two conversation sides. The data were only from land line handsets.

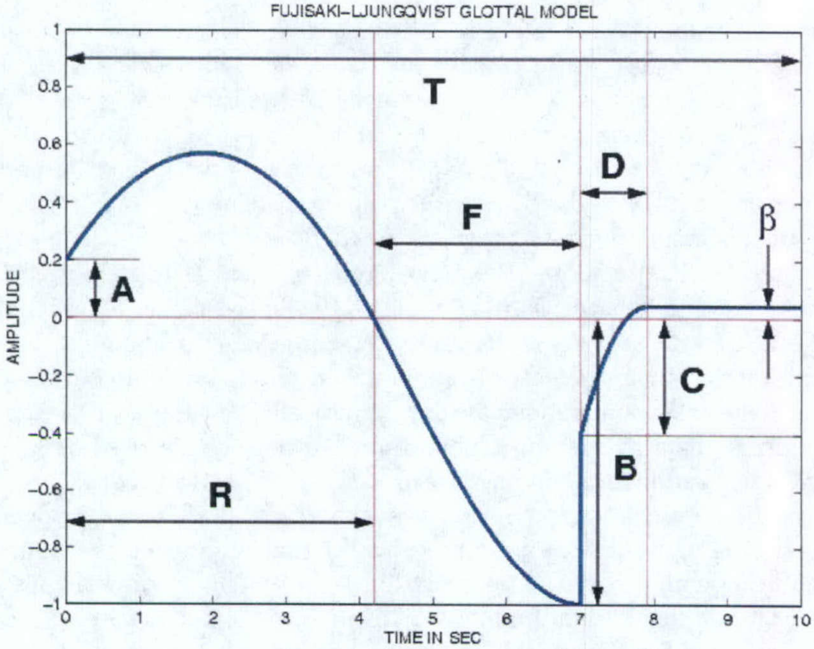


Figure 1: The Fujisaki-Ljungqvist glottal model.

3.1.1 Glottal Model System

This section briefly discusses the glottal model (GM) system [4]. The GM used was a modification of one originally proposed by Fujisaki and Ljungqvist [5].

Figure 1 shows a plot of the Fujisaki-Ljungqvist GM (FLGM) [5], a piecewise polynomial model of the effective voice source. The equation for the model over a pitch period is:

$$g(t) = \begin{cases} A - \frac{2A + R\alpha}{R}t + \frac{A + R\alpha}{R^2}t^2 & \text{for } t \in (0, R], \\ \alpha\,(t-R) + \frac{3B - 2F\alpha}{F^2}\,(t-R)^2 - \frac{2B - F\alpha}{F^3}\,(t-R)^3 & \text{for } t \in (R, W], \\ C - \frac{2\,(C-\beta)}{D}\,(t-W) + \frac{C-\beta}{D^2}\,(t-W)^2 & \text{for } t \in (W, W+D], \\ \beta & \text{for } t \in (W+D, T], \end{cases}$$

where

$$\begin{aligned} \alpha &= \frac{-4AR - 6FB}{F^2 - 2R^2}; \\ \beta &= \frac{CD}{D - 3\,(T-W)}; \end{aligned}$$

T is the pitch period; $W = R + F$ is the duration of the open phase; and A, B, C, D, F, and R are parameters of the model.

Note that the FLGM equations as given in [5] contain two errors. First, the denominator of the t^2 term for $t \in (0, R]$ is given as R in [5] but should be R^2 as given here. Second, the first term in the numerator of α involving $4AR$ is positive in [5] but should be negative as given here. The formula for α comes from the fact that the integral of the model over a pitch period should be zero so as not to introduce a long term trend. Choosing α as given here ensures that the FLGM has the desired integral property.

When estimating the parameters of a GM, one should ensure that the parameter values yield physically meaningful models. However, with the original FLGM, it can be difficult to enforce proper constraints. The timing parameters R, F, and D are not defined in terms of T; thus, as T increases, so do the allowable ranges for R, F, and D. For a physically meaningful model, one cannot have R greater than or equal to T—likewise, for F and D. Further, one cannot have $R+F$, $R+D$, $F+D$, or $R+F+D$ greater than or equal to T. The parameter A must be nonnegative but is theoretically unbounded above, and the parameter B must be negative but is theoretically unbounded below. Practically speaking, one often assumes that $|A| < |B|$, but this means that the upper bound on A varies as B varies.

For the work described here, a modified form of the FLGM was used. There were two main issues that drove the modifications. First, it was important to allow for more easily enforcing proper parameter constraints in the parameter estimation algorithm. Second, the FLGM is defined from glottal opening to glottal opening; however, reliably finding the instants of glottal opening is much more difficult than finding the instants of glottal closure. Thus, the modified FLGM (MFLGM) was defined from glottal closure to glottal closure.

Figure 2 shows a plot of the MFLGM. Note that the plot shows the time as a fraction of the pitch period, T, and that the timing parameters are defined in terms of T. Thus, the time that the glottis is open is qT, giving q as the open quotient (*i.e.*, the fraction of the pitch period over which the glottis is open). One can define p as an "opening" quotient, the fraction of the open phase during which the glottal flow is increasing; thus, pqT is equivalent to R in the FLGM. The m, q, and r parameters are all bounded by zero and one, making it easy to enforce proper constraints on them. The parameter p has tighter bounds. It must be less than one, and it must be greater than or equal to 0.5. The lower bound of 0.5 ensures that the glottal model has the proper skew (*i.e.*, the time duration over which the glottal flow increases is greater than or equal to the duration over which the glottal flow to decreases). The equivalent of the B parameter is constrained to be -1 in the MFLGM, and the gain is handled elsewhere in the parameter estimation process. Finally, the u parameter, while theoretically unbounded above, generally can be constrained in practice to be less than one, given that $B=-1$. Let

$$\begin{aligned}\tilde{\alpha} &= \frac{6(1-p)-4up}{\left[(1-p)^2-2p^2\right]qT},\\ \tilde{\beta} &= \frac{mr}{3-r},\end{aligned}$$

and $\gamma = (1-q)T + pqT$, then the MFLGM is

$$\tilde{g}(t) = \begin{cases} -m + \dfrac{2(m+\tilde{\beta})}{r(1-q)T}t - \dfrac{m+\tilde{\beta}}{r^2(1-q)^2T^2}t^2 & \text{for } t \in (0, r(1-q)T], \\[2ex] \tilde{\beta} & \text{for } t \in (r(1-q)T, (1-q)T], \\[2ex] u - \dfrac{2u+pqT\tilde{\alpha}}{pqT}\left(t-(1-q)T\right) + \dfrac{u+pqT\tilde{\alpha}}{p^2q^2T^2}\left(t-(1-q)T\right)^2 & \text{for } t \in ((1-q)T, \gamma], \\[2ex] \tilde{\alpha}\left(t-\gamma\right) - \dfrac{3+2(1-p)qT\tilde{\alpha}}{(1-p)^2q^2T^2}\left(t-\gamma\right)^2 + \dfrac{2+(1-p)qT\tilde{\alpha}}{(1-p)^3q^3T^3}\left(t-\gamma\right)^3 & \text{for } t \in (\gamma, T]. \end{cases}$$

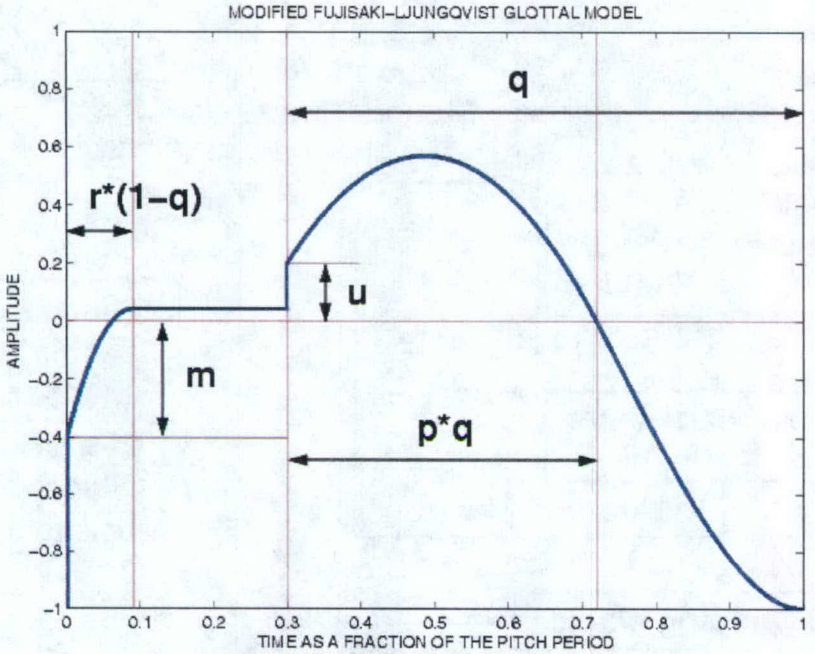


Figure 2: The modified Fujisaki-Ljungqvist glottal model.

The parameter estimation procedure consists of three phases as shown in Figure 3. The first phase determines the instants of glottal closure and opening, which are used to determine closed phases for analysis. The second phase performs a smoothed closed-phase (CP) analysis of the speech signal followed by inverse filtering in order to obtain a residual signal, and the third phase estimates the GM parameters from the residual signal.

In Figure 3, the blocks to the left of the smoothed CP analysis block are designed to get the instants of glottal closure and opening. To find the glottal closures, one performs linear prediction (LP) analysis on the speech signal every 10 msec and inverse filters the speech signal with the resulting LP models to get a residual signal. The data used in the NIST evaluation was sampled at 8 kHz, so the LP analysis used a model order of 10. In parallel with this step, one estimates F0 and the probability of voicing every 10 msec. For the work described here, the Entropic *get_f0* command was used to estimate F0 and the probability of voicing.

Next, one uses a peak picker to determine the quasi-periodic instants of maximum excitation in the residual, which are assumed to correspond to glottal closures. The Entropic *epochs* command was used to perform this task in early experiments; however, it often chose peaks with spacings that did not correspond to the F0 estimates from the *get_f0* program. Instead, another peak picker was used. The peak picker used the probability of voicing to determine the segments of the residual over which to find peaks; it did not pick peaks in unvoiced segments. The peak picker used F0 to help determine peak spacing, working from strong peaks in the middle of each voiced segment both forward and backward to the ends of the segment and choosing strong peaks with spacings that roughly corresponded to the F0 estimates from *get_f0*. It used additional passes to attempt to fix pitch doubling and halving. Note that depending on the glottal model used, the "closure" marks may just signify the points of maximum excitation rather than strict closure points (hence the quotes around the word "Closure" in Figure 3).

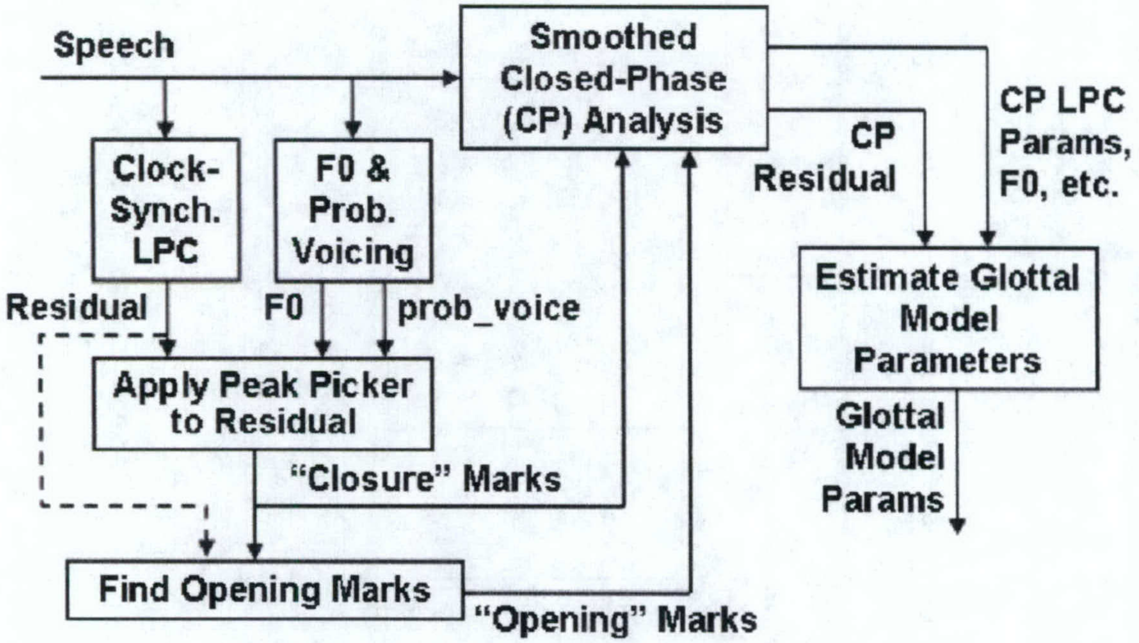


Figure 3: Block diagram of the glottal model parameter estimation procedure.

The last step of this phase is to find the instants of glottal opening. One might conceive of using the residual and the glottal closures (and perhaps other information) to determine the instants of glottal opening; however, as previously mentioned, reliably determining instants of glottal opening is difficult due to the fact that the abruptness of glottal closure leads to a much more pronounced effect on the effective voiced source compared to that due to the more gradual glottal opening. Instead of using the residual and a more complicated procedure to estimate the instants of glottal opening, it was decided to fix the glottal opening points by using a nominal closed quotient of 30% for each pitch period (*i.e.,* an opening was "declared" to occur 30% of the way into each pitch period). The fixed closed quotient means that the next phase of the estimation procedure is not doing *strict* CP analysis and is the reason for the quotes around the word "Opening" in Figure 3. The value of 30% was chosen as a compromise between using enough data to estimate an LP model and choosing a value small enough so as to be doing CP analysis in at least an approximate sense, but additional research should be done to more fully explore the impact of this trade-off. For some pitch periods of high F0, a 30% closed phase did not provide enough data points to estimate the parameters of the LP model. In these cases, the CP analysis extended the closed phase just enough to get enough data to estimate the LP model. Note that even though a nominal closed quotient of 30% is used in the CP analysis, the open quotient parameter, q, is allowed to vary in the glottal model estimation procedure.

One further item to note is that the F0 value that is fed into the "Estimate Glottal Model Parameters" block comes from the spacing between adjacent "closure" marks, not from the original estimate of the *get_f0* command.

Smoothed CP analysis is a generalization of standard CP analysis. In standard CP analysis, for a given pitch period, one forms a correlation matrix, R_0, and a correlation vector, r_0, according to the covariance method of LP [6] and solves the equation $R_0 x = r_0$ for the vector of LP polynomial coefficients, x. Smoothed CP analysis adds the correlation matrices from adjacent pitch periods as well as the correlation vectors prior to computing the LP polynomial. Let R_i be the i^{th} correlation matrix, r_i be the i^{th} correlation vector, and x be the vector of the LP polynomial coefficients for the 0^{th} pitch period. Compute $R_S = R_{-1} + R_0 + R_1$ and $r_S = r_{-1} + r_0 + r_1$, then solve $R_S x = r_S$ for x. In sliding to the next pitch period, the current R_0 becomes R_{-1}, the current r_0 becomes r_{-1}, the current R_1 becomes R_0, and so on. Thus, only one new correlation matrix and one new correlation vector need to be computed each pitch period, just as in standard CP analysis. For the NIST evaluation, one additional pitch period from both the left and the right was used; although, one might want to try a factor of two or more for speech from females due to their higher pitch. Also, for the NIST evaluation, the analysis used an LP order of ten and a preemphasis factor of 0.97. Prior to inverse filtering the speech signal, the stability of the LP polynomials was checked and any unstable roots were reflected about the unit circle.

When estimating the GM parameters from the CP residual, the first step is to normalize the gain of the CP residual (to account for setting $B = -1$ in the MFLGM). For each pitch period, fit a line through the absolute value of the two closure points that define it. This line gives a y-intercept parameter, g_0, and a slope parameter, g_1. Divide each point of the CP residual for the pitch period by its corresponding point on the line. This procedure yields the gain-normalized CP residual.

Typically in glottal modeling, one estimates the model parameters by minimizing the squared error between the model and the residual in the time-domain. However, the low-pass filtering performed prior to sampling (to reduce aliasing) and the distortion of the telephone bandwidth can lead to errors in the parameter estimates. In this work, a frequency-domain metric was used—namely, the squared error between the sine and cosine terms of the harmonic coefficients of the gain-normalized CP residual and those of the MFLGM. (For each pitch period, the spacing between the closure points defining the pitch period was used to determine the value of F0 to be used in computing the harmonics.) Let $g_d(k)$ be the k^{th} data point in the gain-normalized CP residual for a pitch period, then $g_d(k)$ is modeled as:

$$g_d(k) = a_{d,0} + \sum_{n=1}^{N} a_{d,n} \cos(\omega_0 nkT_s) + b_{d,n} \sin(\omega_0 nkT_s)$$

for $k = \{0, \ldots, K-1\}$, where $\omega_0 = 2\pi\mathrm{F0}$, T_s is the sampling time, K is the number of data points in the pitch period, and N is the number of harmonics that lie within the bandwidth of the speech signal. The harmonic coefficients of the data, $a_{d,0}$, $a_{d,n}$, and $b_{d,n}$ for $n = 1, \ldots N$, are estimated by solving a linear matrix equation using the pseudoinverse. The Fourier series expansion of the MFLGM yields the harmonic coefficients of the model in terms of m, p, q, r, and u. The Fourier series of the GM is found as

$$\begin{aligned}
\tilde{g}(t) &= a_{m,0} + \sum_{n=1}^{\infty} a_{m,n} \cos(\omega_0 nt) + b_{m,n} \sin(\omega_0 nt), \\
a_{m,0} &= \frac{1}{T_0} \int_0^{T_0} \tilde{g}(t) dt, \\
a_{m,n} &= \frac{2}{T_0} \int_0^{T_0} \tilde{g}(t) \cos(\omega_0 nt) dt, \\
b_{m,n} &= \frac{2}{T_0} \int_0^{T_0} \tilde{g}(t) \sin(\omega_0 nt) dt,
\end{aligned}$$

where $T_0 = 1/\mathrm{F0}$. (Note that $a_{m,0} = 0$ due to the previously mentioned integral property of the GM.) It was found that the speaker recognition performance improved by multiplying the MFLGM harmonics by an overall gain term, G, so this gain was included in the final metric used for the NIST evaluation. The metric was

$$\text{metric} = \frac{1}{N}\sum_{n=1}^{N}\left(\sqrt{a_{d,n}^2 + b_{d,n}^2} - G\sqrt{a_{m,n}^2 + b_{m,n}^2}\right)^2.$$

The chosen metric is nonlinear in the parameters of the GM. While one could employ iterative estimation techniques, these can converge to local minima. A simpler but more computationally intensive procedure was used here. Offline, use the Fourier series expansion of the MFLGM to build a large linked codebook of (m, p, q, r, u) 5-tuples and their corresponding harmonic coefficient vectors. For a given pitch period and a given harmonic vector from the codebook, find the gain that minimizes the squared error between the gain-weighted model harmonics and those of the gain-normalized CP residual. The best harmonic vector for the pitch period in terms of the metric yields the optimal model parameters for the pitch period, $(m^*, p^*, q^*, r^*, u^*)$, as well as the optimal gain, G^*, and the optimal metric, metric*.

The MFLGM used for the NIST evaluation used $u = 0$ to reduce the computational load in the parameter estimation procedure. Experiments in allowing u to vary from zero to 0.6 in steps of 0.1 yielded better fit errors between the model and the data than did setting $u = 0$, but allowing u to vary did not yield any benefit in speaker recognition performance and required considerably more computation than did setting $u = 0$. Thus, u was set to zero and was not used as a feature for speaker recognition. For each pitch period, the set of features used for speaker recognition was $\{m^*, p^*, q^*, r^*, G^*, \text{metric}^*, g_0, g_1, \mathrm{F0}\}$.

Using the codebook in one serial process is very time consuming; however, the calculations required for each codebook vector are independent of those for the other vectors. Thus, one can break the full codebook into pieces and use them in separate parallel processes to speed up the overall parameter estimation. Each process returns the optimal set of parameters for its codebook piece, and a final process aggregates the separate results into a final optimal parameter set for each pitch period. For the NIST evaluation, the full codebook used 270,000+ vectors on the following grid: $0 \le m \le 1$ in steps of 0.05, $0.5 \le p < 1$ in steps of 0.02, $0.4 \le q < 1$ in steps of 0.02, $0 < r \le 0.3$ in steps of 0.02, and $u = 0$.

The glottal model features were then used in the MIT-LL GMM system and a gender-independent T-norm was applied to normalize the scores from the GMM system. For the 10-second and 30-second training conditions, each T-norm model was built from 30 seconds of data. For the one-side training condition, each T-norm model was built using a side from a single five-minute conversation.

3.1.2 Linear Prediction-Based Cepstral Coefficient System

The LPCC system calculated 16 cepstral coefficients (excluding the 0^{th} cepstral coefficient) from the LP parameters [6] derived from the CP analysis used to find the glottal model parameters. Cepstral mean subtraction was applied to the cepstral coefficients, and the feature set included the deltas of the features.

These features were then used in the GMM system and a gender-independent T-norm was applied. For 10-second and 30-second training conditions, each T-norm model was built from 30 seconds of data. For the 1-side training condition, each T-norm model was built using a side from a single five-minute conversation.

3.1.3 Formant Center Frequencies, Formant Bandwidths, and F0 System

The FMBWF0 system was similar to that of [7]. First, F0 and the probability of voicing were determined every 10 msec using the Entropic Signal Processing System (ESPS) *get_f0* command. Next, the first three formant center frequencies (F1–F3) and the first three formant bandwidths (B1–B3) were determined from Wavesurfer 1.6.2 (and Snack 2.2.2) from KTH.[3] Each F0 value was converted to log scale. Each formant center frequency and bandwidth value was converted to radians.

Extracted frames had (1) to be declared to be speech by the *xtalkN* SAD, (2) to be voiced; (3) to have F0 $<$ 250 Hz; and (4) to have F1 $\neq$ 500 Hz, F2 $\neq$ 1500 Hz, and F3 $\neq$ 2500 Hz. Condition (3) was imposed because the pitch extractor was found to output pitch-doubled frames at times, while condition (4) was imposed to eliminate frames where the formant tracker failed.

These features were then used in the GMM system and a gender-independent T-norm was applied. For 10-second and 30-second training conditions, each T-norm model was built from 30 seconds of data. For the 1-, 3-, 8-, and 16-side training conditions, each T-norm model was built using a side from a single five-minute conversation.

3.1.4 Mel-Frequency Cepstral Coefficient System

Mel-frequency cepstral coefficients (MFCCs) were computed using the MIT-LL GMM system [2]. Nineteen MFCCs were calculated from the speech waveform and output every 10 msec. RASTA filtering was applied to the MFCCs and deltas were then calculated. Only frames labeled as speech by the *xtalkN* energy-based SAD were used.

These features were then used in the GMM system and a gender-independent T-norm was applied. For 10-second and 30-second training conditions, each T-norm model was built from 30 seconds of data. For the 1-, 3-, 8-, and 16-side training conditions, each T-norm model was built using a side from a single five-minute conversation.

3.1.5 Phoneme-Specific Mel-Frequency Cepstral Coefficient System

The PS-MFCC system was similar to the system described in [8] that used phoneme-only adaptation. The main difference between this system and the one in [8] had to do with how the phoneme labels were determined. The system described here used Sonic (Version 2.0-beta1) [9,10] *run as a phoneme recognizer, not as a speech recognizer*, whereas the system of [8] used the phoneme labels from speech recognition transcripts provided by Stanford Research Institute (SRI) for the NIST 2003 Extended Data Task. The acoustic models for Sonic were provided by Prof. Brian Pellom (the original developer of Sonic) of the University of Colorado at Boulder and were trained using land line data from Switchboard. A trigram "language" model with phonemes in place of words was also built from Switchboard.

Background, target, and T-norm models were built as follows. MFCCs were computed exactly as in the standard MFCC system of Section 3.1.4 (*i.e.,* using the MIT-LL GMM system), and each feature vector was associated with a phoneme label as output by Sonic. The phonemes used were from the following set: {AE, N, AY, AH, M, AX, S, Y, IY, L, OW, IH, K, EY, R, EH, AA, W}. A separate background GMM (of either 1024 or 2048 mixtures, depending on the phoneme) was built for each phoneme using data from only Switchboard II (*i.e.,* no cellular data was used). For each target and T-norm speaker, a GMM was built for each phoneme, which

[3] Available at: http://www.speech.kth.se/wavesurfer and http://www.speech.kth.se/snack

was adapted from the background model for that phoneme. Thus, each target and T-norm "model" was actually a collection of GMMs with each GMM labeled with a specific phoneme. For the 1-, 3-, 8-, and 16-side training conditions, each T-norm "model" was built using two conversation sides of data from Switchboard II with phoneme transcripts generated by Sonic.

Scoring an utterance against a claimant "model" proceeded as follows. First, each feature vector of the utterance was assigned a phoneme label by Sonic. Next, all the vectors for a given phoneme were scored against the claimant's GMM built for that phoneme, and a phoneme-dependent T-norm was applied. Finally, the scores for each phoneme (after the phoneme-dependent T-norm had been applied) were combined with a perceptron neural net that had been trained using the MIT-LL LNKnet package.[4] The neural net used no hidden layers, and the output nonlinearity was a standard sigmoid. The neural net had been trained using data from the NIST 2003 Extended Data Task.

3.2 Language Modeling Systems

Two of the component systems were based on language modeling. The first was based on modeling word bigrams, while the second system was based on modeling phoneme bigrams.

3.2.1 Word-based Language Model System

The CMU-Cambridge Language Modeling Toolkit[5] (Version 2.05) formed the basis of this system. A speech recognizer from BBN was used to generate transcripts, and these transcripts were provided to the evaluation participants by NIST. The words from the transcripts were assembled into pseudo sentences, where a pause greater than one second between words defined a sentence break. Using no sentence breaks, where each conversation side became one sentence, yielded worse performance than using pseudo sentence breaks.

Bigram language models were trained with the following parameters set in the toolkit: top 20,000 words, Witten-Bell discounting, and zero cut-offs. Target models were trained by concatenating all the sentences for each of the conversations allowed for each model, while the background model was built in a similar way, but with all the sentences from all the files that made up the background model.

To compute a score using the word-based language modeling (WLM) system, the sentences from a test file were tested against a claimant model and the background model. The score for a given test file and claimant model pair was computed as follows. Let B_C be the set of bigrams in the claimant model, C; B_B be the set of bigrams in the background model; and B_T be the set of bigrams in a test file, T. Let $B_{TCB} = B_T \cap B_C \cap B_B$, and let N_{TCB} be the number of bigrams in B_{TCB}. Let $\Pr(b, C)$ be the probability of bigram b in model C and $\Pr(b, B)$ be the probability of bigram b in the background model. The score for T against the claimant model C was computed as:

$$\text{score}(T, C) = \frac{1}{N_{TCB}} \sum_{b \in B_{TCB}} \log(\Pr(b, C)) - \log(\Pr(b, B)).$$

Thus, unknown or non-matching bigrams were ignored.

[4] Available at: http://www.ll.mit.edu/IST/lnknet

[5] Available at: http://svr-www.eng.cam.ac.uk/ prc14/toolkit.html

One final step was taken with the inclusion of a gender-independent T-norm. 50 male and 50 female models were built using two conversation sides of data from Switchboard II[6] with transcripts generated by a BBN speech recognizer.[7]

3.2.2 Phoneme-based Language Model System

The phoneme-based language model (PLM) system was similar to the WLM system, except that it used bigrams of phonemes rather than bigrams of words. The phonemes were determined by using Sonic [9, 10] as a phoneme recognizer rather than as a speech recognizer. Unlike the word language model system, no T-norm was applied.

3.3 Unsupervised Adaptation

For each system submitted using this paradigm, the same thresholds were used for both the true/false decisions and for updating the speaker models. The thresholds used for 3-side training were interpolated from the thresholds found from the 2- and 4-conversation side training of the NIST 2003 Extended Data Evaluation. The adaptation procedure consisted of evaluating a test file against a claimant model, and if the resultant score was above the threshold, then the model was updated for use the next time. A model was updated by adding the current test utterance to the current training data used to build a model and rebuilding the model. Using this strategy yielded no improvement in performance over the standard (nonadaptive) mode of operation.

3.4 HMM-Based Speech Activity Detector

For the segmentation tasks, a SAD based on HMMs was used rather than just the energy-based SAD discussed in Section 3.1. After the evaluation results were submitted, the HMM-based SAD was tried for the other tasks and found to perform better than the energy-only SAD. This section describes the HMM-based SAD. Results of using the HMM-based SAD with some of the GMM-based systems are shown in Section 5.

Figure 4 shows a block diagram of the HMM-based SAD. The feature extractor computed 19 MFCCs and deltas (with no cepstral mean subtraction or RASTA filtering) using the Hidden Markov Model Toolkit (HTK).[8] A two-state speech/non-speech HMM was built with HTK using 80 mixtures per state. The HMM was trained on 100 Switchboard II speech files using label files provided by SRI. The *xtalkN* energy-based detector refined the output from the HMM-based detection; thus, segments were declared to be speech only if they satisfied both detectors. The noise floor for *xtalkN* was set using the average frame energy from the top ten non-speech segments from the HMM-based detection. Finally, the post-processing removed any speech segments of less than 20 msec in duration.

3.5 Segmentation

This section describes the segmentation-based system, which used the MFCC feature set described in Section 3.1.4 and the HMM-based SAD described in Section 3.4. Each speech file was

[6] See http://www.ldc.upenn.edu

[7] Note that the recognizer used to generate transcripts for Switchboard II was most likely not the same as that used to generate transcripts for the 2004 Evaluation data.

[8] Available from: http://htk.eng.cam.ac.uk

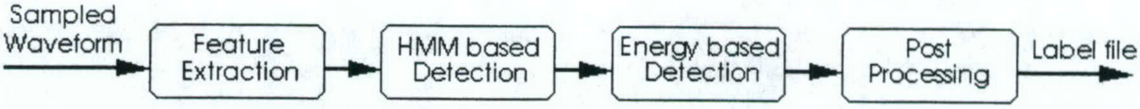


Figure 4: Block diagram of HMM-based speech activity detector.

segmented into speaker homogeneous regions. This was accomplished by using the HMM-based SAD to define utterance boundaries and then clustering similar speech segments.

An agglomerative clustering method was used to cluster similar speech segments. In order to determine which segments should be clustered together, a GMM system was used. A 64-mixture GMM was trained using all of the vectors classified as speech, and then the weights of this model were adapted to fit the characteristics of each speech segment (thus creating a separate model for each speech segment). In each stage of the clustering, the feature vectors for each speech segment were scored against all of the models and the highest scoring feature vector/model pair were merged. This process was repeated until three sets of segments were left (presumably, a set of segments for each of the two speakers and a set of "garbage" segments).

For the three-conversation training conditions, the same clustering method was used across the three speech files to determine the common speaker in the three files. For the one-conversation test conditions, each of the three sets of segments was tested against the claimant model and the highest score was taken as the overall score. No T-norm was applied for the actual evaluation, but after the evaluation, T-norm was added for the one-conversation testing cases.

4 SUBMITTED SYSTEMS

The submitted systems were as outlined in Tables 2 and 3. Systems 1–3 were simple linear weightings of the scores from componenet systems. System 5 was used for the segmentation tasks. Systems 6 and 7 used perceptron neural nets (with no hidden layers and an output sigmoid nonlinearity) trained on the NIST 2003 Extended Data Task. Systems 4, 8, 9, and 10 were component systems submitted for comparison purposes. Systems 8–10 were also submitted with unsupervised adaptation.

Table 2: Submitted Systems in Terms of Component Systems

#	System Description
1	0.7*score(MFCC with T-norm)+0.3*score(FMBWF0 system with T-norm)
2	0.7*score(MFCC with T-norm)+0.15*score(FMBWF0 system with T-norm) +0.15*score(LPCC system with T-norm)
3	0.7*score(MFCC with T-norm)+0.1*score(FMBWF0 system with T-norm) +0.1*score(LPCC system with T-norm)+0.1*score(GM system with T-norm)
4	MFCC system with T-norm
5	MFCC system without T-norm but using HMM SAD and clustering
6	Fusion (with LNKnet) of scores from the MFCC system with T-norm, the FMBWF0 system with T-norm, and MFCC systems with T-norm for each of the following phonemes from the Sonic speech recognizer: {AE, N, AY, AH, M, AX, S, Y, IY, L, OW, IH, K, EY, R, EH, AA, W}
7	Fusion (with LNKnet) of scores from MFCC systems with T-norm for each of the following phonemes from the Sonic speech recognizer: {AE, N, AY, AH, M, AX, S, Y, IY, L, OW, IH, K, EY, R, EH, AA, W}
8	FMBWF0 system with T-norm
9	Language modeling applied to phonemes output by Sonic (without T-norm)
10	Language modeling applied to the words from the BBN transcripts with T-norm

Table 3: Submitted Systems Versus Training and Testing Condition

Training	Testing Condition			
Condition	10 sec	30 sec	1 side	1 conv
10 sec	1, 2, 3, 4	1, 2, 3, 4	1, 2, 4	5
30 sec	1, 2, 3, 4	1, 2, 3, 4	1, 2, 4	5
1 side	1, 2, 3, 4	1, 2, 3, 4	1, 2, 4, 6, 7, 8 (UA), 9 (UA), 10 (UA)	5
3 sides	1, 4	1, 4	1, 4, 6, 7, 8 (UA), 9 (UA), 10 (UA)	5
8 sides	1, 4	1, 4	1, 4, 6, 7, 8 (UA), 9 (UA), 10 (UA)	5
16 sides	1, 4	1, 4	1, 4, 6, 7, 8 (UA), 9 (UA), 10 (UA)	5
3 conv	5	5	5	5

5 RESULTS

This section presents some results comparing the performance of the various systems on data from both the NIST 2004 and 2003 evaluations. The performance results are in the form of detection error tradeoff (DET) plots [11], which are plots of miss probability versus false alarm probability with both axes using a normal deviate scale.

5.1 The NIST 2003 Limited Data Task and the NIST 2004 One-Side Training and One-Side Testing Task

Figure 5 shows a DET plot for the FMBWF0, GM, LPCC, and MFCC systems for both the NIST 2003 Limited Data Task and the comparable task from the 2004 Evaluation—namely, the one-side training and one-side testing task. It is clear from this plot that the 2004 task is considerably more difficult than the 2003 task for all four systems. This difficulty arises from at least three factors. First, the 2004 data had much greater channel variability compared to the 2003 data, which was almost entirely cellular data (mostly CDMA and some GSM). Second, the speech activity detector didn't perform as well as it could have on the 2004 data; this will be discussed more in Section 5.2. Finally, the 2004 data allowed languages other than English, so a speaker model could be in one language while a testing file from that speaker could be in another language, thereby generating some phonetic mismatch.

In addition to using DET plots, system performance is often summarized in terms of equal error rate (EER), the value of the false alarm probability and the miss probability when the two are equal. Table 4 shows the EERs for the same four component systems on the two tasks as well as the absolute and percentage increases in EER relative to the EERs for the 2003 evaluation. From this table, one can see that the MFCC system performed the best in terms of EER in both the 2003 and 2004 evaluations; however, the FMBWF0 system was the most consistent across the two evaluations as evidenced by the fact that it yielded the lowest absolute and percentage increases in EER. While the MFCC system performed better than the LPCC system in terms of EER and absolute increase in EER, the two systems performed comparably in terms of the percentage increase in EER. Finally, the GM system was the least consistent, but performed comparably to the FMBWF0 and LPCC systems for the 2003 evaluation. The considerably better performance on the 2003 evaluation by the GM system is presumably due to the greater channel homogeneity in the 2003 evaluation relative to that of the 2004 evaluation.

5.2 Effect of the HMM SAD

After the official system scores were submitted, experiments were performed to assess the effect of using the HMM SAD with various component systems. Figures 6 and 7 show the effects of using the HMM SAD versus the energy-only SAD for the FMBWF0 system, the MFCC system, and the combination of the FMBWF0 and MFCC systems (*i.e.*, submitted system 1) for one-side testing with one-side training and eight-sides training, respectively. Use of the HMM SAD improved the performance of the FMBWF0 system only a little, but it improved the performance of the MFCC system quite a bit, resulting in a comensurate improvement in the combination of the two systems. The FMBWF0 system required that frames not only be labeled as speech but also that they be labeled as voiced, whereas the MFCC system used all frames labeled as speech. Thus, one can see that the HMM SAD made the biggest difference in the speech/non-speech decisions for the unvoiced frames.

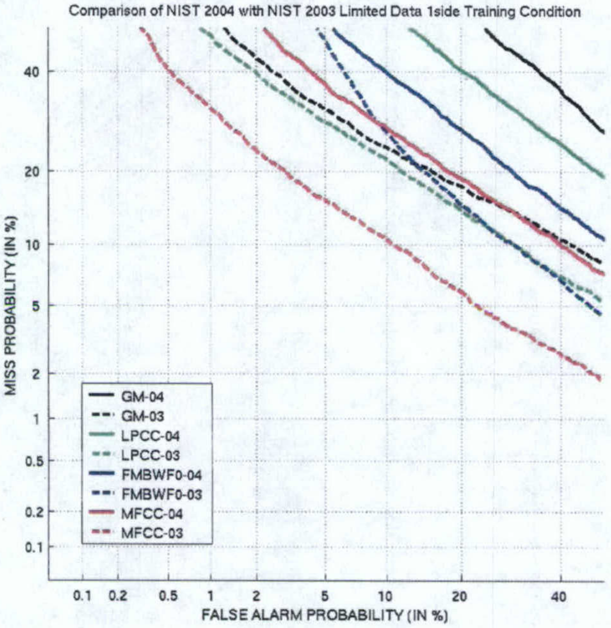


Figure 5: DET plot comparing the NIST 2003 Limited Data Task with the NIST 2004 one-side training and one-side testing task.

Table 4: Equal Error Rates (EERs) for Four Component Systems for 2003 and 2004

Component System	EER for NIST 2003 Limited Data Task	EER for NIST 2004 One-Side Train & One-Side Test Task	Absolute Increase in EER Relative to 2003 EER	Percentage Increase in EER Relative to 2003 EER
FMBWF0 with T-norm	17.2%	24.2%	7.0%	41%
MFCC with T-norm	10.3%	19.3%	9.0%	87%
LPCC with T-norm	16.3%	30.9%	14.6%	90%
GM with T-norm	18.2%	37.9%	19.7%	108%

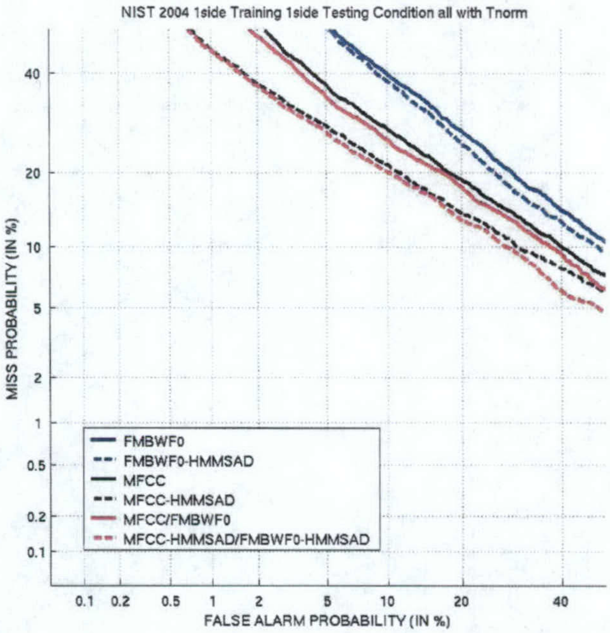


Figure 6: DET plot for the NIST 2004 evaluation one-side training and one-side testing condition showing the effect of the HMM SAD.

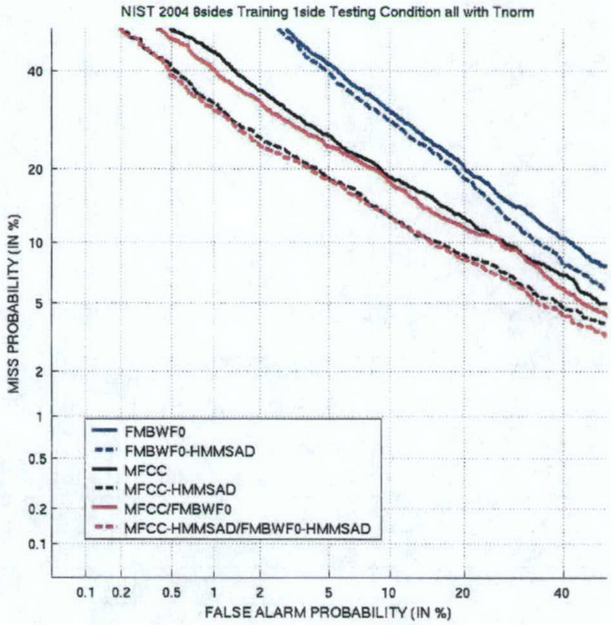


Figure 7: DET plot for the NIST 2004 evaluation eight-sides training and one-side testing condition showing the effect of the HMM SAD.

5.3 Effect of Training Length for the NIST 2003 and 2004 Evaluations

Figures 8 and 9 show DET plots comparing the performance of the WLM, FMBWF0, MFCC, and PS-MFCC systems for one-side testing with one-side training and eight-sides training, respectively, for the NIST 2003 Extended Data and NIST 2004 evaluations. Again, one can see that the performance of all of the systems is better for the 2003 evaluation than for the 2004 evaluation. Regardless of the evaluation year or the number of training conversations, the PS-MFCC system performed the best, followed by the MFCC, FMBWF0, and WLM systems in that order. In general, the PS-MFCC system requires only one-quarter to one-third the training data to perform comparably to the standard MFCC system. The FMBWF0 system requires about eight times the training data to perform comparably to the MFCC system, while the WLM system requires about eight times the training data to perform comparably to the FMBWF0 system.

5.4 Effect of Combining System Scores for the NIST 2004 Evaluation

Figures 10 and 11 show the effect of combining component system scores for the NIST 2004 evaluation with one-side testing and one- and eight-sides training, respectively. For both training lengths, the combination of the FMBWF0, MFCC, and PS-MFCC systems outperforms any of the individual WLM, FMBWF0, MFCC, and PS-MFCC component systems. Further combining the WLM system score with the scores of the other three component systems provides almost no benefit for the single-side training condition, but provides as much benefit over the combination of the other three systems as the combination of the other three component systems provides over the PS-MFCC system.

5.5 One-Conversation Testing in the NIST 2004 Evaluation

Figure 12 shows the DET plot for the one-conversation testing conditions of the NIST 2004 evaluation for the MFCC system with and without T-norm, which was integrated after the official evaluation submission. For every type of training level, T-norm provided a benefit. A particularly interesting comparison is that between the three-side training condition and the three-conversation training condition, in which one must determine what segments contain the common speaker across the three training files. One can see that the three-conversation condition results in considerably worse performance relative to the three-side training condition. In fact, the three-conversation training condition performs even worse than the 30-second training condition. On the other hand, comparing the MFCC system in the one-conversation testing and one-side training case with the MFCC system in the one-side testing and one-side training case of Figure 10, one can see that the performance is comparable. Thus, there appears to be room for significant improvement in the three-conversation training case.

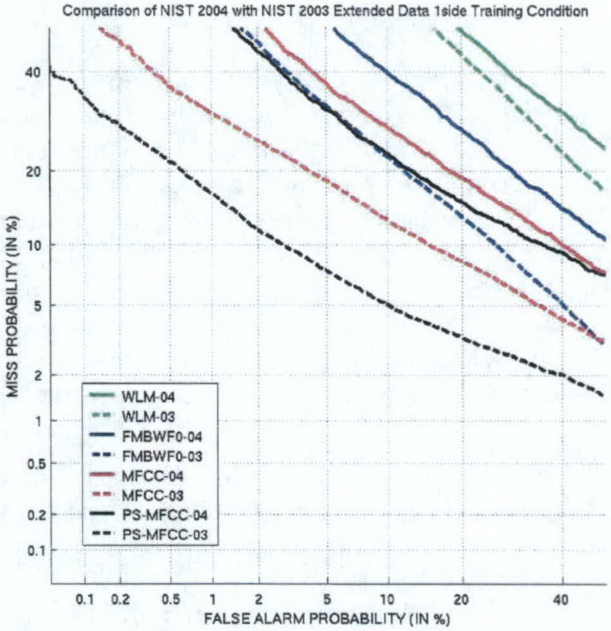


Figure 8: DET plot comparing NIST 2004 evaluation with the NIST 2003 Extended Data evaluation with one-side training and one-side testing.

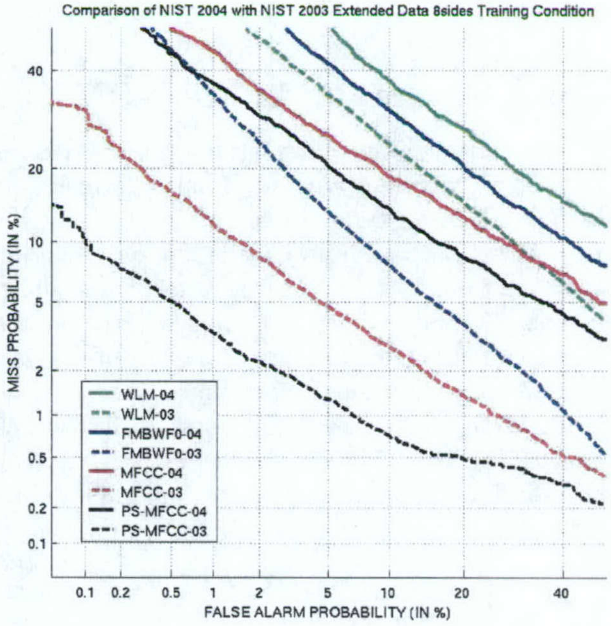


Figure 9: DET plot comparing NIST 2004 evaluation with the NIST 2003 Extended Data evaluation with eight-sides training and one-side testing.

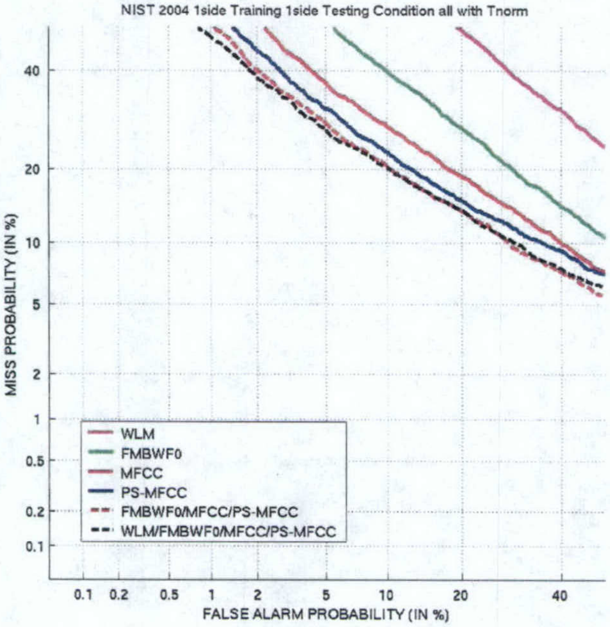


Figure 10: DET plot showing the effect of combining system scores for the NIST 2004 evaluation with one-side training and one-side testing.

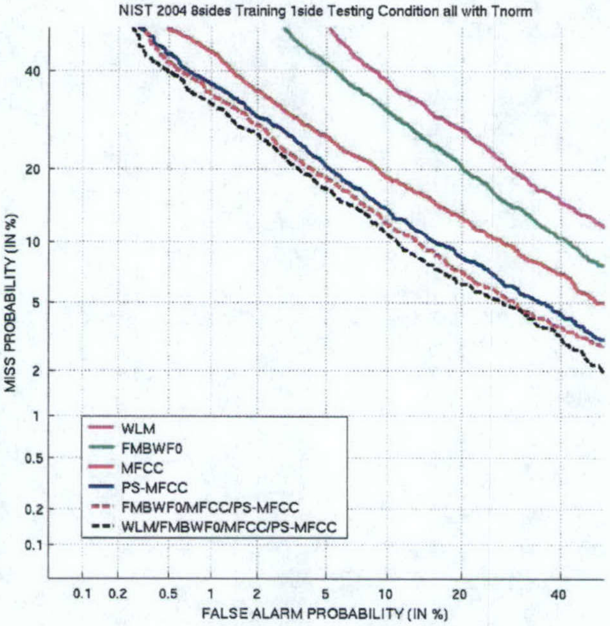


Figure 11: DET plot showing the effect of combining system scores for the NIST 2004 evaluation with eight-sides training and one-side testing.

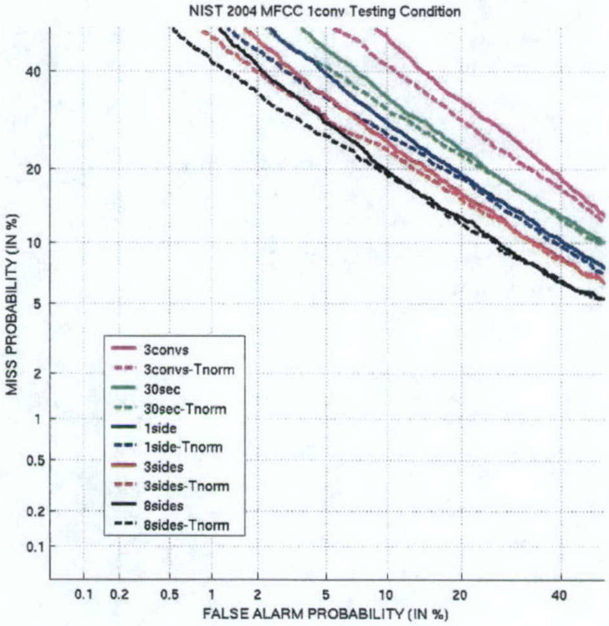


Figure 12: DET plot for one-conversation testing conditions with and without T-norm.

6 CONCLUSIONS

As can be seen from the performance of the various systems on the NIST 2004 data relative to their performance on the NIST 2003 data, channel robustness continues to be an issue that needs further work. This is especially the case for the GM and LPCC systems. Language and accent issues also require further work, especially for the PS-MFCC system. A multilingual phoneme recognizer and/or speech recognizer might help in this regard. For unsupervised adaptation to be useful, additional work needs to be done to determine when to update a speaker's model. For the segmentation tasks, there appears to be room for significant improvement in the three-conversation training conditions. Finally, combining system scores improves the performance compared to that of single systems.

REFERENCES

[1] NIST, *The NIST Year 2004 Speaker Recognition Evaluation Plan*, Version 1a, 29 January 2004. Available at: `http://www.nist.gov/speech/tests/spk/2004/index.htm`.

[2] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, nos. 13, pp. 19–41, 2000.

[3] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing,* vol. 10, nos. 1-3, pp. 42–54, 2000.

[4] R. Slyh, E. Hansen, and T. Anderson, "Glottal modeling and closed-phase analysis for speaker recognition," in *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop,* (Toledo, Spain), May–June 2004.

[5] H. Fujisaki and M. Ljungqvist, "Proposal and evaluation of models for the glottal source waveform," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* (Tokyo, Japan), vol. 3, pp. 1605–1608, April 1986.

[6] J. Deller, J. Proakis, and J. Hansen, *Discrete-Time Processing of Speech Signals,* MacMillan: New York, 1993.

[7] E. Hansen, R. Slyh, and T. Anderson, "Formant and F0 features for speaker recognition," in *Proceedings of A Speaker Odyssey: The Speaker Recognition Workshop,* (Chania, Crete, Greece), June 2001.

[8] E. Hansen, R. Slyh, and T. Anderson, "Speaker recognition using phoneme-specific GMMs," in *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop,* (Toledo, Spain), May–June 2004.

[9] B. Pellom and K. Hacioglu, "Recent improvements in the CU SONIC ASR system for noisy speech: The SPINE task", in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* (Hong Kong), April 2003.

[10] B. Pellom, *SONIC: The University of Colorado Continuous Speech Recognizer,* University of Colorado, Technical Report TR-CSLR-2001-01, (Boulder, Colorado), March 2001.

[11] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance", in *Proceedings of EuroSpeech '97,* (Rhodes, Greece), September 1997.